Binary Classification of ClinVar Clinical Classification Confliction

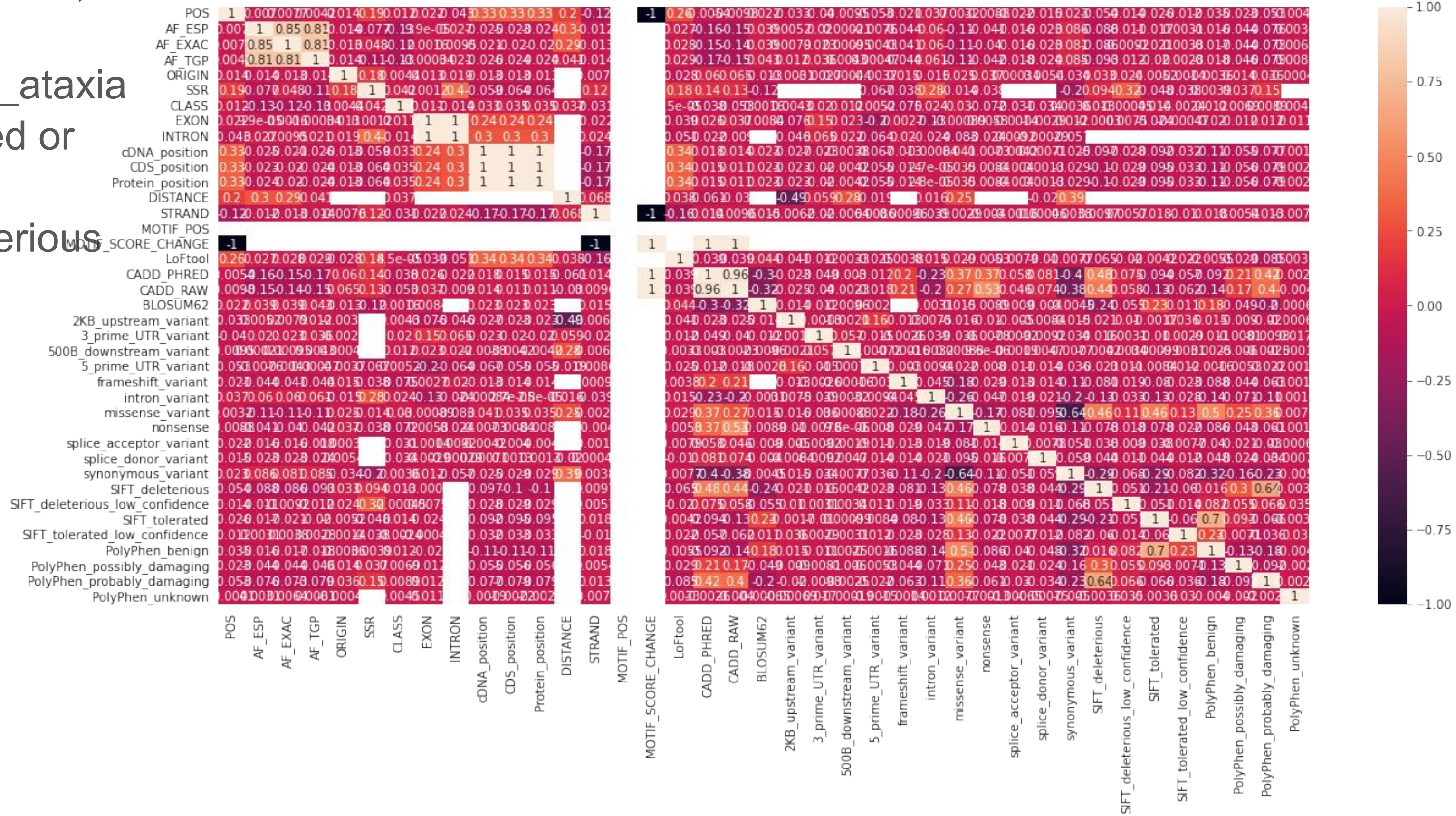
Anna Lee-Hassett
MACS 33002
Prof Zhao Wang
Spring 2022

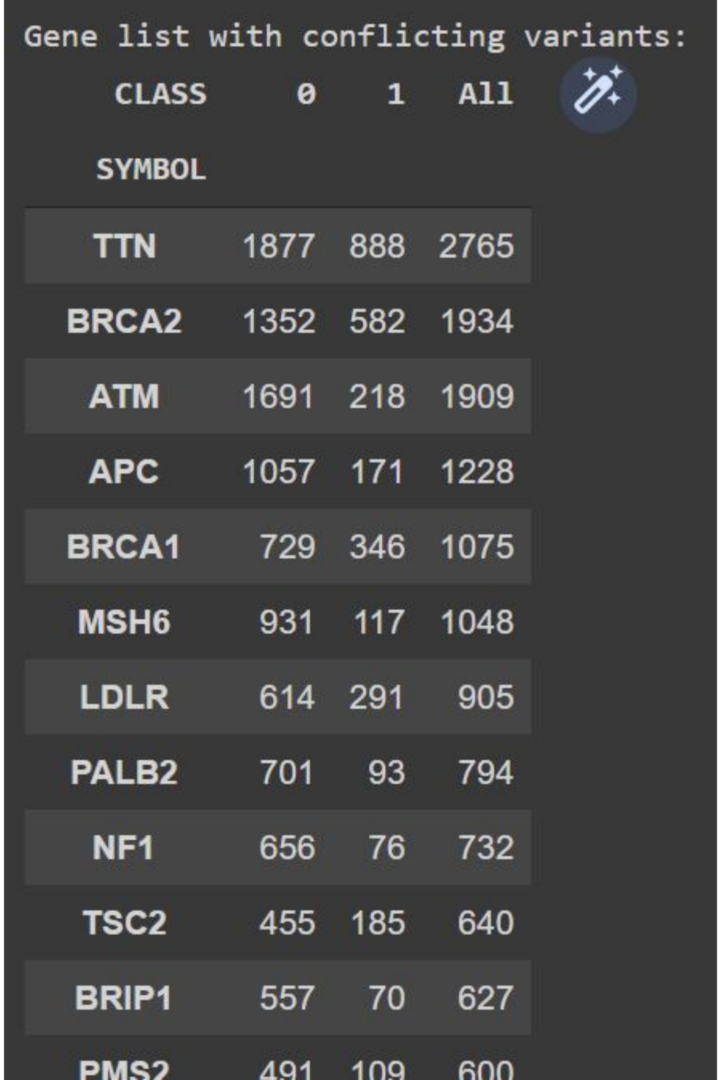
Data

- CHROM = chromsome the variant is located on
- POS = variant's position on chromosome
- REF = reference allele (non mutated)
- ALT = alternate allele (mutated)
- AF_ESP = allele frequencies from GO-ESP
- AF_EXAC = allele freq from ExAC
- AF_TGP = allele freq from 1000 genomes proj
- CLNDISDB = description of disease associated with the variant. Stored as a tag-value pair of disease database name and identifier. Within one variable entry, different diseases are separated by "|", different databases within the same disease separated by ",".
- CLNDISDBINCL = for included variant, the above (all values = nan)
- CLNDN = ClinVar's prefered disease name for the concept specified by disease identifiers in CLNDISDB.
- CLNDNINCL = for included variant, the above (all values = nan)
- CLNHGVS = a valid HGVS expression based on top-level genomic sequences (assembled chromosomes, mitochondrion, or alternate loci or patches).
- CLNSIGINCL = clinical significance for a haplotype or genotype that includes this variant (all values = nan)
- CLNVI = variant type (SNV, deletion, other, etc.)
- CLNVI = clinical sources stored as tag-value pairs of database:variant identifier (most nan)
- MC = comma separated list of molecular consequences in the form of sequence ontology ID|molecular_consequence
- ORIGIN = allele origin. 0 - unknown; 1 - germline; 2 - somatic; 4 - inherited; 8 - paternal; 16 - maternal; 32 - de-novo; 64 - biparental; 128 - uniparental; 256 - not-tested; 512 - tested-inconclusive; 1073741824 - other
- SSR = variant suspect reason codes. One of more of the following: 0 - unspecified, 1 - Paralog, 2 - byEST, 4 - oldAlign, 8 - Para_EST, 16 - 1kg_failed, 1024 - other
- CLASS = target variable (see below)
- Allele = variant allele used to calculate the conseqeunce
- Consequence = type of variant consequence, such as "splice_donor_variant," "stop_lost", "missense_variant", or "intron_variant"
- IMPACT = subjective classification of the severity of variant consequence, based on SNPEff stored as LOW, MODERATE, etc.
- SYMBOL = Gene name
- Feature_type = Transcript, RegulatoryFeature, MotifFeature
- Feature = Ensembl stable ID of feature
- BIOTYPE = all protein_coding?
- EXON = exon number (out of total number) (14% [null])
- INTRON = intron number (out of total number) (86% [null])
- cDNA_position = relative position of base pair in cDNA sequence
- CDS_position = relative position of base pair in coding sequence
- Protein_position = relative pos of amino acid in protein
- Amino_acids = affected amino acids, [null] if variant doesn't affect protein-coding seq
- Codons = alternative codons with variant base in uppercase (ex cGg/cAg)
- DISTANCE = shortest distance from variant to transcript (100% [null]?)
- STRAND = forward (+), reverse (-)
- BAM_EDIT = success/failure of edit using a BAM file (51% [null], 49% "OK")
- SIFT = SIFT (Sorting Intolerant from Tolerant alg) prediction and/or score. Predicts effect of coding vars on protein function. Mostly [null], some "deleterious", some other.
- PolyPhen = PolyPhen prediction and/or score
- MOTIF_NAME = source and identifier of transcription factor binding profile (TFBP) aligned at this position (100% [null])
- MOTIF_POS = relative position of variation in algined TFBP (100% [null])
- HIGH_INF_POS = flag indicating if variant falls in high information position of a TFBP (100% [null])
- MOTIF_SCORE_CHANGE = diffreence in motif score of regerence and variant seq for TFBP (100% [null])
- LoFtool = LOF tolerance score for LOF variants
- CADD_PHRED = Phred-scaled CADD (combined annotation dependent depletion) score. Predicts varient effect. (Phred score estimates the probability a NT base was sequenced incorrectly. Higher q-score = more confidence.)
- CADD_RAW = score of deleteriousness (harm) of variants
- BLOSUM62 = assignment of alignment score to subsituted amino acids caused by variants

Encoding ←→ data visualization

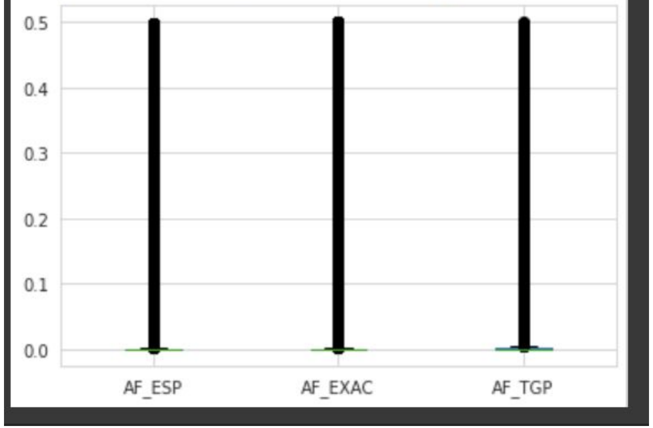
- 8028/8030
- MedGen:CN169374, OMIM:607454
- Spinocerebellar_ataxia_21|not_provided or not_specified
- Tolerated, deleterious
- NaN

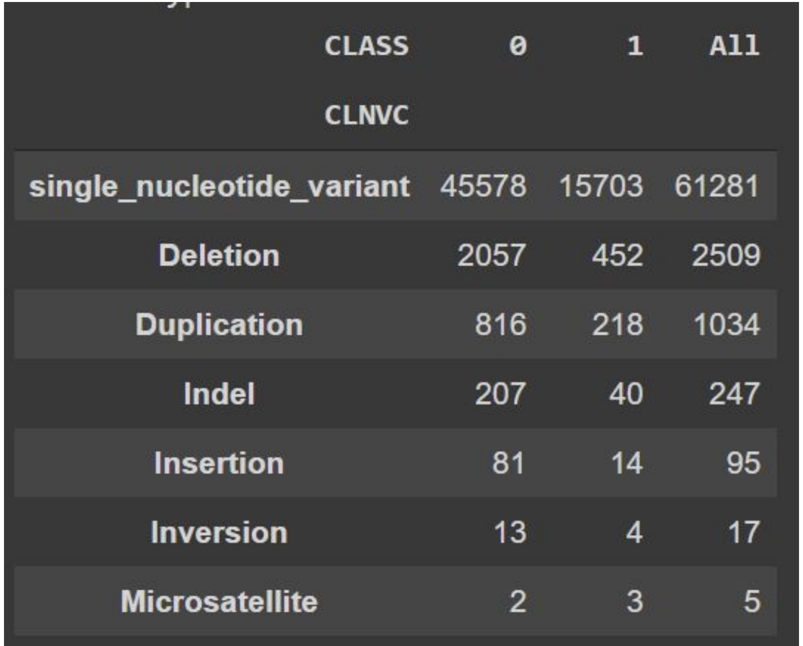




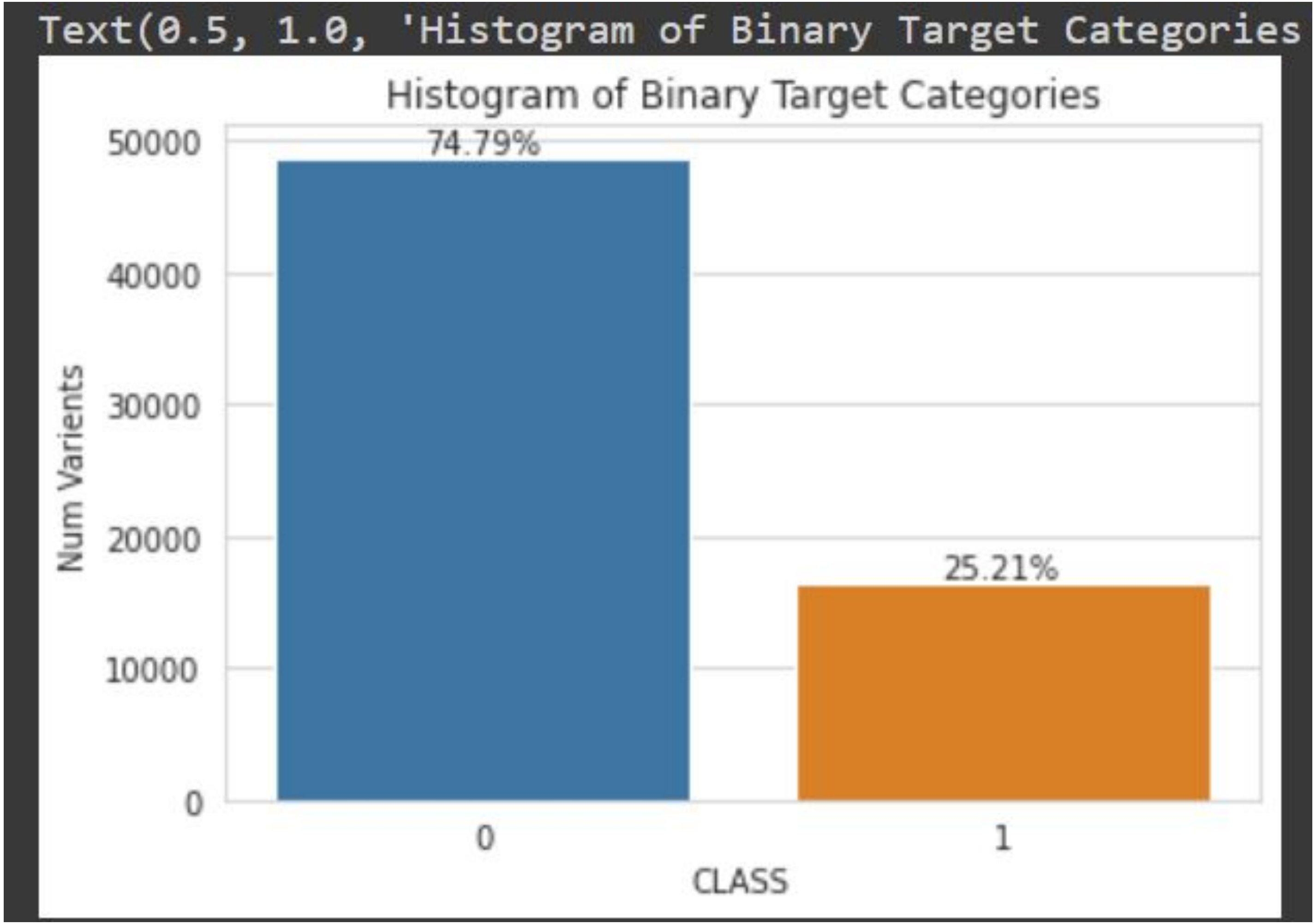
Gene list with conflicting variants:

CLASS	0	1	All
SYMBOL			
TTN	1877	888	2765
BRCA2	1352	582	1934
ATM	1691	218	1909
APC	1057	171	1228
BRCA1	729	346	1075
MSH6	931	117	1048
LDLR	614	291	905
PALB2	701	93	794
NF1	656	76	732
TSC2	455	185	640
BRIP1	557	70	627
PMS2	491	109	600





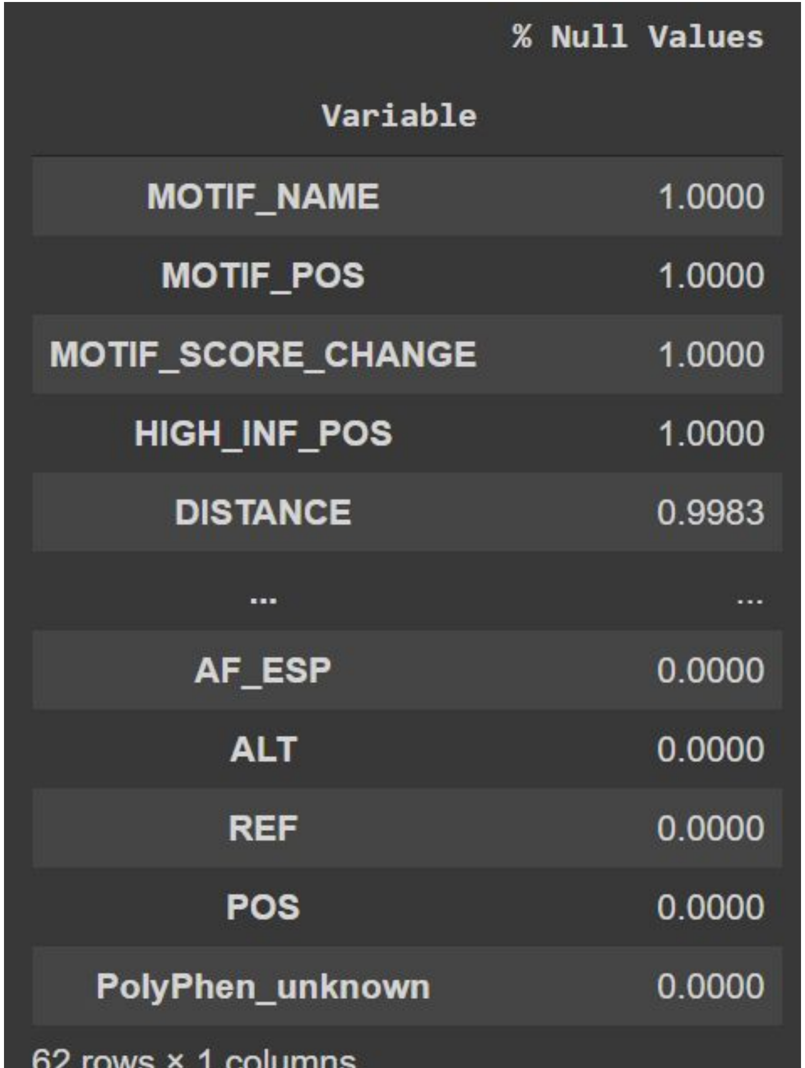
CLASS	0	1	All
CLNVC			
single_nucleotide_variant	45578	15703	61281
Deletion	2057	452	2509
Duplication	816	218	1034
Indel	207	40	247
Insertion	81	14	95
Inversion	13	4	17
Microsatellite	2	3	5



```
Text(0.5, 1.0, 'Histogram of Binary Target Categories
```

Data removal

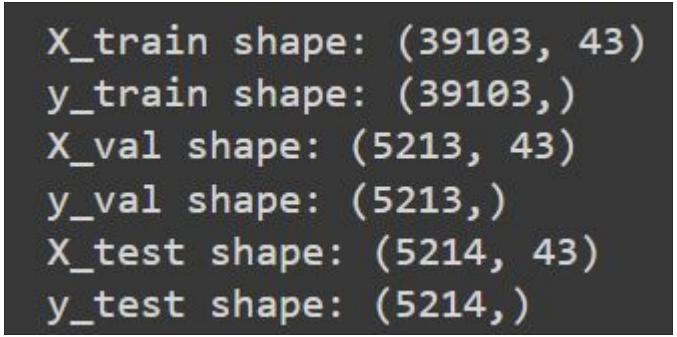
- >99% missing
- Without gene name (‘Symbol’ = NaN)
- Entirely unique values (CLNHGVS)
- Entirely identical values
    - ‘Feature_type’, ‘BIOTYPE’
- Multilinearity correlation/repetitive data
    - 'AF_TGP', 'AF_EXAC', 'cDNA_position', 'CDS_position', and 'CADD_RAW',
- Duplicates (none)



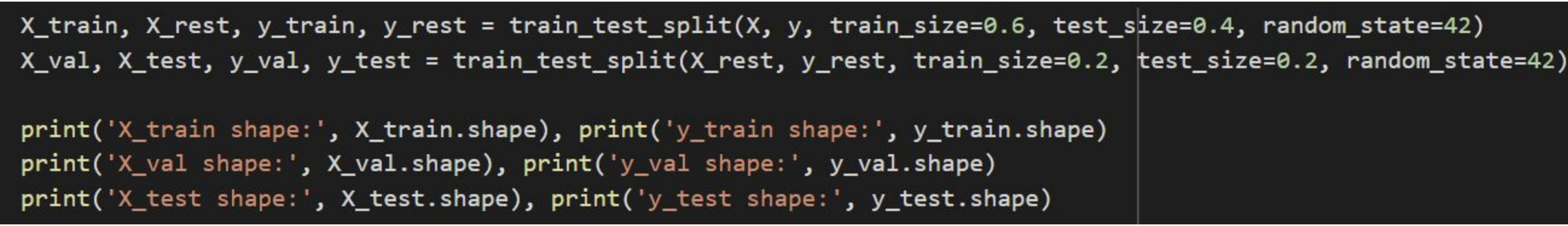
	% Null Values
Variable	
MOTIF_NAME	1.0000
MOTIF_POS	1.0000
MOTIF_SCORE_CHANGE	1.0000
HIGH_INF_POS	1.0000
DISTANCE	0.9983
...	...
AF_ESP	0.0000
ALT	0.0000
REF	0.0000
POS	0.0000
PolyPhen_unknown	0.0000

62 rows × 1 columns

Train/Val/Test split



```
X_train shape: (39103, 43)
y_train shape: (39103,)
X_val shape: (5213, 43)
y_val shape: (5213,)
X_test shape: (5214, 43)
y_test shape: (5214,)
```

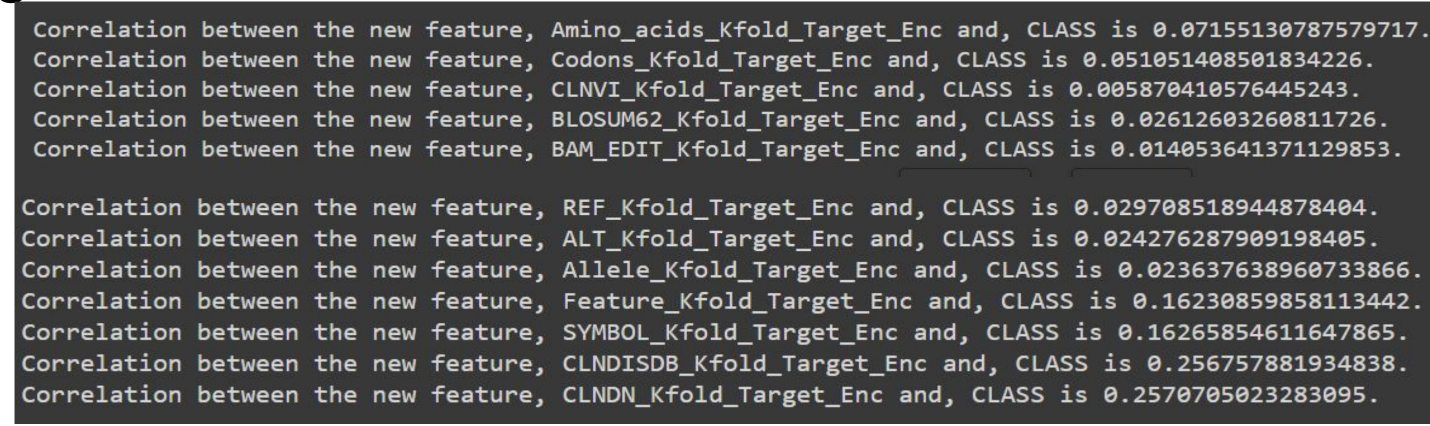


```
X_train, X_rest, y_train, y_rest = train_test_split(X, y, train_size=0.6, test_size=0.4, random_state=42)
X_val, X_test, y_val, y_test = train_test_split(X_rest, y_rest, train_size=0.2, test_size=0.2, random_state=42)

print('X_train shape:', X_train.shape), print('y_train shape:', y_train.shape)
print('X_val shape:', X_val.shape), print('y_val shape:', y_val.shape)
print('X_test shape:', X_test.shape), print('y_test shape:', y_test.shape)
```

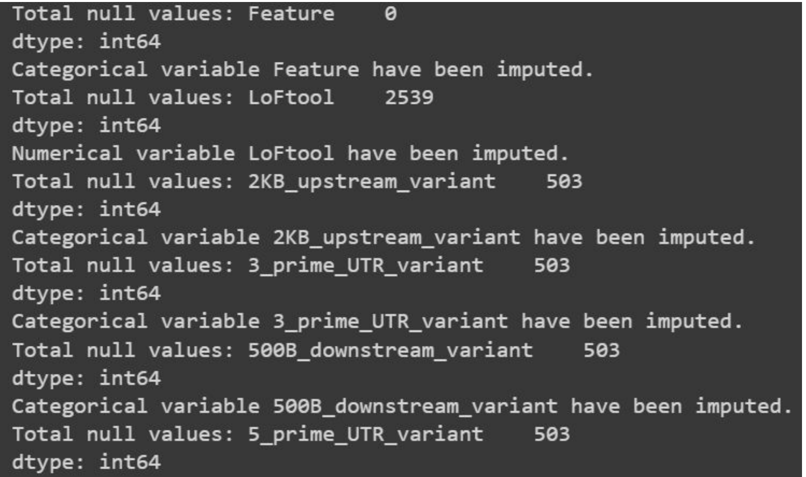
More Processing

- Sparse column removal
- Outlier detection
- Encoding (kfold)
- Imputation (average-based)
- Scaling

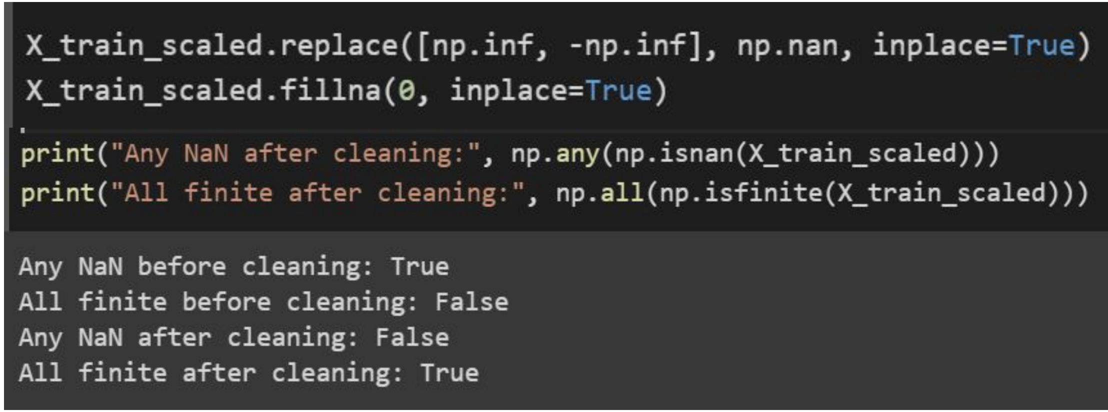


```
Correlation between the new feature, Amino_acids_Kfold_Target_Enc and, CLASS is 0.07155130787579717.
Correlation between the new feature, Codons_Kfold_Target_Enc and, CLASS is 0.05105140850183422б.
Correlation between the new feature, CLNVI_Kfold_Target_Enc and, CLASS is 0.005870410576445243.
Correlation between the new feature, BLOSUM62_Kfold_Target_Enc and, CLASS is 0.02612603260811726.
Correlation between the new feature, BAM_EDIT_Kfold_Target_Enc and, CLASS is 0.014053641371129853.

Correlation between the new feature, REF_Kfold_Target_Enc and, CLASS is 0.029708518944878404.
Correlation between the new feature, ALT_Kfold_Target_Enc and, CLASS is 0.024276287909198405.
Correlation between the new feature, Allele_Kfold_Target_Enc and, CLASS is 0.023637638960733866.
Correlation between the new feature, Feature_Kfold_Target_Enc and, CLASS is 0.16230859858113442.
Correlation between the new feature, SYMBOL_Kfold_Target_Enc and, CLASS is 0.16265854611647865.
Correlation between the new feature, CLNDISDB_Kfold_Target_Enc and, CLASS is 0.256757881934838.
Correlation between the new feature, CLNDN_Kfold_Target_Enc and, CLASS is 0.257070502328309б.
```



```
Total null values: Feature    0
dtype: int64
Categorical variable Feature have been imputed.
Total null values: LoFtool    2539
dtype: int64
Numerical variable LoFtool have been imputed.
Total null values: 2KB_upstream_variant    503
dtype: int64
Categorical variable 2KB_upstream_variant have been imputed.
Total null values: 3_prime_UTR_variant    503
dtype: int64
Categorical variable 3_prime_UTR_variant have been imputed.
Total null values: 500B_downstream_variant    503
dtype: int64
Categorical variable 500B_downstream_variant have been imputed.
Total null values: 5_prime_UTR_variant    503
dtype: int64
```



```
X_train_scaled.replace([np.inf, -np.inf], np.nan, inplace=True)
X_train_scaled.fillna(0, inplace=True)
```

```
print("Any NaN after cleaning:", np.any(np.isnan(X_train_scaled)))
print("All finite after cleaning:", np.all(np.isfinite(X_train_scaled)))
```

```
Any NaN before cleaning: True
All finite before cleaning: False
Any NaN after cleaning: False
All finite after cleaning: True
```

Model fit/transform

- LogReg
- Gradient Boost
  - Principle: multiple weak learners (usually DTs) → strong classifier
  - Functions by iteratively adjusting model by adding more components and calculating loss
- XG Boost
  - Similar to gradient boost, but trees can have varying terminal nodes. Includes a randomization parameter that reduces correlation between ensemble trees, which aids in strength of classifier.

Errors/Conclusion/Improvement

- Improved streamlining of preprocessing
- Convergence Errors